# Specific Communication Network Measure Distribution Estimation

Daniel P. Baller, Joshua Lospinoso
United States Military Academy
West Point, NY 10996

*Abstract*— **A new method is proposed to estimate the probability distribution of specific communication network measures. Real world communication networks are dynamic and vary based on an underlying social network, thus reliably estimating network measures is challenging. Two individuals that are socially connected may communicate several times one day, and not at all on another, yet their basic relationship remains unchanged. In this situation, estimates of network measures, such as density, degree centrality and others may be severely affected by the occurrence or absence of observed communication ties between individuals.**

**The communication network of a group of mid-career Army officers is modeled from empirical data using the network probability matrix (NPM) proposed by McCulloh and Lospinoso (2007). The NPM provides a framework to model a communication network by estimating the edge probabilities between two individuals in a network. This framework can model a specific social group regardless of their network topology: random, small-world, scale-free, cellular, etc. Monte Carlo simulation is used with the NPM to generate 100,000 instances of the communication network. A statistical distribution is fit to the density measure. Using this probability distribution, statistically significant changes in density can be detected.**

*Index Terms*—**NPM, Network Probability Matrix, Social Network, Density, Distribution**

## I. Introduction

Various techniques are used in the network science community for the simulation of networks. These frameworks typically are based on the topology and structure of the network i.e. triads, dyads and cliques. However theses techniques do not always take into account all of the factors that contribute to the dyadic relationship between agents. In a network an agent may not care that there is a triad between 3 other agents or that certain agents in the network have dyadic ties. The agent is primarily concerned with his or her own dyadic relationships leading to an underlying dynamic equilibrium in the network.

This dynamic equilibrium involves an underlying edge probability structure that contains a probability that each agent will communicate with every other agent in the network.

This probability structure remains constant in the network independent of observations at a single instance in time. In a single observation the appearance of a tie does not indicate that a relationship exists as the communication may have been made in error. Conversely the lack of communication between two agents in a single observation does not indicate the lack of a relationship as an agent is not consistently communicating with every agent he has a relationship with at all times. While the appearance or lack of communication does not indicate that the relationship between two agents exists, the communication at a single observation relies on the underlying probability that the agents will communicate.

The network probability matrix (NPM) proposed by McCulloh and Lospinoso (2007) posits that networks can be simulated based on the underlying probability structure of the dynamic equilibrium. This framework estimates the edge probabilities between each combination of two individuals in a network. Probability estimation can range from a proportion of communications in a series of observations or be estimated from more complex distributions depending on the amount and type of data present. This structure can then be used to simulate a variety of network topologies: random, small-world, scale free, cellular, ect.

The edge probability structure of the underlying dynamic equilibrium remains constant in the network while the network is at a stable state. However, it may shift as shocks to the network take place. Using Monte Carlo simulation, the underlying distributions of network measures can be determined while the network is in its dynamic equilibrium. These underlying distributions can be used in change detection and allow us to statically predict shocks to the network and may be an indicator to determine when significant changes occur.

# Report Documentation Page

*Form Approved*
*OMB No. 0704-0188*

| | | |
|---|---|---|
| 1. REPORT DATE<br>**JUN 2008** | 2. REPORT TYPE | 3. DATES COVERED<br>**00-00-2008 to 00-00-2008** |

| | |
|---|---|
| 4. TITLE AND SUBTITLE<br>**Specific Communication Network Measure Distribution Estimation** | 5a. CONTRACT NUMBER |
| | 5b. GRANT NUMBER |
| | 5c. PROGRAM ELEMENT NUMBER |
| 6. AUTHOR(S) | 5d. PROJECT NUMBER |
| | 5e. TASK NUMBER |
| | 5f. WORK UNIT NUMBER |
| 7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)<br>**United States Military Academy,West Point,NY,10996** | 8. PERFORMING ORGANIZATION REPORT NUMBER |
| 9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES) | 10. SPONSOR/MONITOR'S ACRONYM(S) |
| | 11. SPONSOR/MONITOR'S REPORT NUMBER(S) |

12. DISTRIBUTION/AVAILABILITY STATEMENT

**Approved for public release; distribution unlimited**

13. SUPPLEMENTARY NOTES

**13th International Command and Control Research and Technology Symposia (ICCRTS 2008), 17-19 Jun 2008, Seattle, WA**

14. ABSTRACT

**A new method is proposed to estimate the probability distribution of specific communication network measures. Real world communication networks are dynamic and vary based on an underlying social network, thus reliably estimating network measures is challenging. Two individuals that are socially connected may communicate several times one day, and not at all on another, yet their basic relationship remains unchanged. In this situation, estimates of network measures, such as density, degree centrality and others may be severely affected by the occurrence or absence of observed communication ties between individuals. The communication network of a group of mid-career Army officers is modeled from empirical data using the network probability matrix (NPM) proposed by McCulloh and Lospinoso (2007). The NPM provides a framework to model a communication network by estimating the edge probabilities between two individuals in a network. This framework can model a specific social group regardless of their network topology: random, small-world, scale-free, cellular, etc. Monte Carlo simulation is used with the NPM to generate 100,000 instances of the communication network. A statistical distribution is fit to the density measure. Using this probability distribution, statistically significant changes in density can be detected.**

15. SUBJECT TERMS

| 16. SECURITY CLASSIFICATION OF: | | | 17. LIMITATION OF ABSTRACT | 18. NUMBER OF PAGES | 19a. NAME OF RESPONSIBLE PERSON |
|---|---|---|---|---|---|
| a. REPORT<br>**unclassified** | b. ABSTRACT<br>**unclassified** | c. THIS PAGE<br>**unclassified** | **Same as Report (SAR)** | **30** | |

## II. Background

Social network analysis is a theoretical framework that examines the relationships between social entities (e.g. people, groups, organizations, beliefs, knowledge, etc.). These objects are known as nodes and their connections are referred to as edges. Not all nodes are connected while some nodes are connected with multiple relationships. This network framework is applicable in a plethora of content areas such as communications, information flow, and group or organizational affiliation (Titchy & Tushman, 1979). Social network analysis relies heavily on graph theory to make predictions about network structure.

### *A. Erdős-Rénia Random graphs*

In 1959 mathematicians Paul Erdős and Alfréd Rénia made revolutionary discoveries in the evolution of random graphs. In their eight papers Erdős and Rénia evaluate the properties of random graphs with *n* vertices and *m* edges. For a random graph *G* containing no edges, at each time step a randomly chosen edge among the $\binom{n}{2}$ possible edges is added to *G*. This graph contains *N* edges and each edge of the $\left(\begin{matrix}\binom{n}{2}\\ N\end{matrix}\right) = C_{n,N}$ possible edges is equally likely. Therefore, once an edge is chosen from the $\binom{n}{2}$ equally likely edges the next edge is chosen among the remaining $\binom{n}{2}-1$ edges and this process is continued so that if *k* edges are fixed, all remaining $\binom{n}{2}-k$ edges have equal probabilities of being chosen (Erdős & Alfréd Rénia, 1960). A general model used to generate random graphs is as follows: "For a given *p*, $0 \leq p \leq 1$, each potential edge of *G* is chosen with probability *p*, independent of other edges. Such a random graph is denoted by $G_{n,p}$ where each edge is determined by flipping a coin, which has probability *p* of coming up heads (Chung & Graham, 1998)." In this model of random graphs each edge has an equal probability of occurring or not occurring within the graph. This random graph model also assumes that all nodes in the graph are present at the beginning and the number of nodes in the network is fixed and remains the same throughout the network's life. Additionally, all nodes in this model are considered equal and are undistinguishable from each other (Barabási & Albert, 1999).

### *B. Statistical tests*

Utilizing Erdos' theory of random graphs as well as the class of uniform distributions associated with these graphs, Holland and Leinheart (1971) developed a variety of statistical tests for the analysis of social networks. Using a uniform distribution these tests spread the total probability mass equally over all possible outcomes, therefore giving an equal probability to the existence of an edge between any two nodes in the network. These statistical tests were used to develop a reference frame or constant benchmark to which observed data could be compared in order to determine how "structured a particular network was, or how far the network deviated from the benchmark (Furst & Wasserman, 1994)."

### *C. Strength of weak ties*

In 1969, Mark Granovetter proposed the strength of weak ties. In Granovetter's social world our close friends are often friends with each other as well, leading to a society of small, fully connected circle of friends who are all connected by strong ties. These small circles of friends are connected through weak ties of acquaintances. In turn, these acquaintances have strong connections within their own circle of friends. The weak ties connecting circles of friends play an imperative role in numerous social activities from finding a job to spreading the latest fad. Close friends who have strong connections are often exposed to the same information, therefore, weak ties are activated to bridge out of our circle of friends and into the outside world (Granovetter, 1973).

### *D. Small World Networks*

Building off of Granovetter's model Duncan Watts and Steven Strogatz (1998) developed the clustering coefficient, dividing the number of links of a node's first order connections by the number of links possible between these first order connections. This clustering coefficient illustrates the interconnectivity of a circle of friends, where a value close to 1 demonstrates all first order connections of a node are connected with each other. Conversely, a value close to 0 shows that a node's first order connections are only connected through that particular node.

### *E. Scale Free Networks*

The clustering coefficient of the Watts-Strogatz small world network model is the first to reconcile clustering with the characteristics of random graphs. According to the Watts-Strogatz model each node is directly connected to each one of its neighbors resulting in a high clustering coefficient. By clustering alone, this model has a high average path length connecting two random nodes. However, by adding only a few random links between nodes of different clusters the average separation between nodes drastically decreases. This model while containing random links between nodes keeps the clustering coefficient relatively unchanged (Watts & Newman, 1999). While the Watts-Strogatz model originally did not add extra links to the graph but randomly rewired some of the links to distant nodes, the addition of random links was proposed by Watts and M. Newman.

According to Albert-László Barabási the random graph theory of Erdős and Rénia was rarely found in the real world. Barabási has found that many real world networks have some nodes that are connected to many nodes and others that are connected to few nodes. His empirical tests showed that the distribution of the number of connections in many networks all followed a power-law distribution. These networks lack the characteristic scale in node connectivity present in random graphs, and therefore, are scale-free (Barabási, 2003). As a result of the number of connections following a power distribution, hubs are created among nodes in the network. A hub is a highly connected node that contains most of the links in the network and creates short paths between any two nodes in the network.

Barabási's model of scale-free networks is constructed around two ideas—growth and preferential attachment. For each time step a new node is added to the network. This illustrates the principal that networks are assembled one node at a time (Barabási & Albert, 1999). Assuming that each new node connects to the existing nodes of the network with two links, the probability that the new node will choose a given node is proportional to the number of links the chosen node has. Therefore, a node with more links has a higher probability of being connected to. This creates a "rich get richer" scenario where nodes with many links continue to grow by collecting new links while newer nodes with lower degrees do not collect as many links (Barabási & Albert, 1999).

Based on a scale-free network model where nodes make connections based completely on preferential attachment the probability that a new node will connect to a node with $k$ links is given by $\frac{k}{\sum_i k_i}$ (Barabási, 2003). This causes the first nodes in the network to develop into hub nodes due to having the longest time to collect links. However it is not always the case that the first nodes in a network develop into the biggest hubs.

### F. *Fitness Model*

In order to account for newer nodes overtaking older nodes as hubs, Barabási constructed the fitness model. Fitness is a nodes ability to collect links relative to every other node in the network and is based on competition in complex systems (Barabási, & Bianconi, 2001). In this new model a node's attractiveness is not determined completely by its number of links, but preferential attachment is driven by the product of the number of links a node has and its fitness. In this model the probability a new node will connect to a node with $k$ links a fitness of $\eta$ is $\frac{k\eta}{\sum_i k_i \eta_i}$ (Barabási, & Bianconi, 2001). Nodes in this model acquire links following the power law distribution of the scale-free model, however, the dynamic exponent, $\beta$, which determines how fast a node acquires new links, is different for each node. This is proportional to a node's fitness, therefore, a node that is twice as fit as another node will obtain nodes twice as fast because its dynamic exponent is twice as large. This "fit-get-rich" model allows nodes to become hubs based on their attractiveness regardless of when they enter the network (Barabási, & Bianconi, 2001).

### G. *Winner Take All Model*

Contrary to the scale-free network model Barabási developed the "winner take all model," which strongly portrays monopolies. The "winner-take-all-model" consists of a single hub and many tiny nodes. This network develops a star topology and nodes do not acquire links following a power law distribution.

### H. *Network Probability Model*

Ian McCulloh and Joshua Lospinoso (2007) proposed a new structure for random communication networks over time, based on empirical data collected on real world networks. This framework, estimates distributions for the time between communication messages, then based on a given time interval the probability of an edge occurring in the network is calculated for every ordered pair of nodes. These probabilities can be constructed through multiple techniques. To derive the probabilities from empirical data collected over several time periods, a proportion of edge occurrences, $e_{ij}$, can be used to estimate probabilities for each cell in the adjacency matrix $a_{ij}$. These probabilities are displayed in a network probability matrix where each cell is the probability that node *i* communicates with node *j*. This framework is capable of generating networks that are similar to scale free networks. Thus, this model can be used to construct any network topology: Erdős-Rénia random, Watts-Strogatz small world, Albert-Barabási scale-free, star, cellular, ect. The NPM model is estimated from empirical data and can be used to simulate realistic observations of relationships in specific organizations.

## III. DATA

This research evaluates the density of a real world network in order to find the underlying distribution of network density. The data was collected from a war fighting simulation in FT Leavenworth, KS in April 2007 by Craig Schreiber and Lieutenant Colonel John Graham. There were 99 participants in the experiment that were monitored over the course of four days. This 99 agent data set was then cut down to 68 agents. These 68 participants served as staff members in the headquarters of the brigade conducting the exercise. The data displays the interactions of agents in a network collected by a self reported communications survey.

## IV. METHOD

Our study explores the distribution of the density measure in simulated networks using the network probability matrix.

Below is an outline of the approach pursued in this study:

### *A. Construction of the Network Probability Matrix*

In order to simulate the network it is necessary for a network probability matrix, (NPM) to be created. Once the datasets were trimmed of the scripted agents, they were symmetrized across the main diagonal in the Organizational Risk Analyzer (ORA) to account for the lack of directionality of communication in the data. Symmetrizing the data also corrects for the informant error of agents not reporting other agents they have communicated with. Next, the datasets were dichotomized to remove the weighting set by the participants. Once the data is dichotomized a one represents communication between two agents and a zero represents the lack of communication between two agents. To construct the NPM all eight data sets were compiled into a single data set consisting of the total number of discrete time periods that each agent communicated with each other agent. This matrix was then divided by the number of discrete time periods to determine the underlying edge probabilities for the network in dynamic equilibrium.

### *B. Simulation Generation*

The NPM was then used as the edge probabilities for a Monte Carlo simulation of the network. In this simulation a random number was generated for each edge. If the random number is less than the edge probability then the edge is added to the graph. This algorithm was used to create 100,000 simulations of the network. Once 100,000 simulations of the network were completed the average density was taken from each simulation to create a dataset of 100,000 network densities.

### *C. Reliability and Consistency*

To analyze the reliability and consistency of our simulations hamming distances were utilized. Using the NPM, 60,000 instances of the network were simulated. The average hamming distance from each empirical data set to every other empirical data set and from each simulated network to each empirical data set. These average hamming distances were then analyzed using a *t*-test.

### *D. Distribution fitting*

The normal distribution was fit to the data using Maximum Likelihood Estimation. An Anderson-Darling goodness of fit test and a comparison of the estimated cumulative distribution function to the data's empirical distribution function indicated a very good fit for the data. In addition, since the density is a linear function of the average node degree, the central limit theorem would suggest that the density is normally distributed, given certain assumptions.

## V. RESULTS

Using the *t*-test it is shown that the simulated networks have a smaller average hamming distance to the empirical data sets than each empirical data set is to each other. This illustrates that the simulated networks give a more reliable and consistent approximation of the underlying distribution. The results of the *t*-test are shown below in Table 1. Where column one is the average hamming distance from each empirical data set to every other empirical data set and column three is the average hamming distance from 60,000 networks simulated with the NPM to each of the empirical data sets. The *p*-value of each test is approximately zero indicating that there is a statistically significant difference between the empirical hamming distances and the simulated hamming distances. Additionally, since

$$\mu_{emperical} - \mu_{simulated} > 0$$

it is shown that the simulated networks are closer to each of the empirical data sets than the empirical data sets are to each other.

Table 1. *t*-test of Average Hamming Distances

| | *M* | *8* | *N* | *60000* | | |
|---|---|---|---|---|---|---|
| **Data Set** | **Mean Hamming Distance to Empirical Networks** | **Standard Deviation of Hamming Distance to Empirical Networks** | **Mean Hamming Distance to Simulated Networks** | **Standard Deviation of Hamming Distance to Simulated Networks** | ***t*-test** | ***p*-value** |
| 1 | 409.286 | 38.560 | 358.094 | 12.775 | 3.755 | 0.00 |
| 2 | 365.857 | 18.298 | 320.097 | 12.739 | 7.073 | 0.00 |
| 3 | 365.857 | 29.043 | 320.164 | 12.793 | 4.450 | 0.00 |
| 4 | 377.857 | 38.247 | 330.674 | 12.773 | 3.489 | 0.00 |
| 5 | 375.286 | 36.100 | 328.377 | 12.796 | 3.675 | 0.00 |
| 6 | 349.857 | 38.159 | 306.078 | 12.785 | 3.245 | 0.00 |
| 7 | 373.8571 | 48.45076 | 327.0728 | 12.82622 | 2.731135 | 0.01 |
| 8 | 362.4286 | 55.63529 | 317.1509 | 12.77754 | 2.301849 | 0.02 |

Once the reliability and consistency of the simulations created using the NPM was established, the distribution of the density could be analyzed. Since density is a linear function of a sample average of a network statistic according to the formula

$$density = \frac{avg \quad \deg ree}{(n-1)},$$

and the sample size, *n*, is greater than 30 the central limit theorem can be used to show that the underlying distribution of network density is the normal distribution, with μ=0.0984374 and σ=0.00396148.
This is also shown in Figure 1.

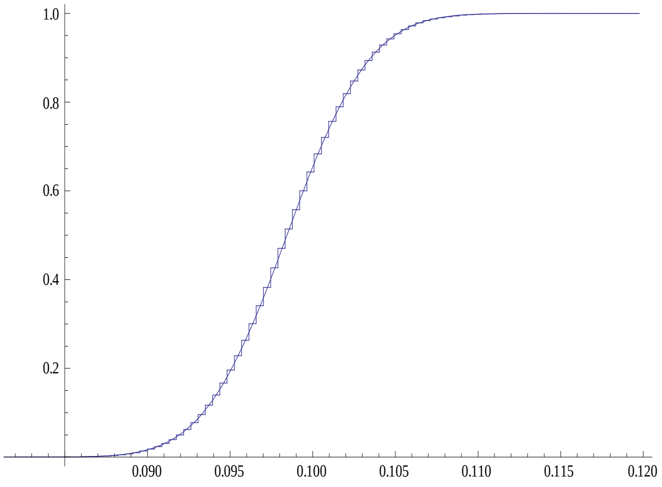


Figure 1. Stepwise Plot of Density Data and CDF of the Normal Distribution

This graph shows the stepwise plot of the 100,000 densities overlaid with the CDF of the normal distribution. The sum of squared error of this model is 9.60609. This small sum of squared error reinforces the model shown above in Figure 1.

## VI. Conclusion

This research validates the use of the NPM for simulating networks based on empirical data. The reliability and consistency of the network simulations provide a strong framework for analysis.

This research can be extended in at least three aspects: assessing the underlying distribution for other network level statistical measures, assessing the underlying distribution for agent level statistical measures, and using these distributions to statistically predict changes and shocks to a network.

## Acknowledgment

We are grateful to MAJ Anthony Johnson and MAJ Ian McCulloh for advising and mentoring in this research effort. We are also indebted to COL Steve Horton for his assistance in creating the Mathematica 6.0 simulations used in this study.

## References

[1] Baller, D. P., Lospinoso, J., McCulloh, I., & Johnson, A. N. (n.d.). Specific Communication Network Measure Distribution Estimation. Unpublished Manuscript.

[2] Baller, D. P., Lospinoso, J., & Johnson, A.N. (2008). *An Empirical Method for the Evaluation of Dynamic Network Simulation Methods.* In Proceedings, The 2008 World Confress in Computer Science Computer Engineering and Applied Computing, Las Vegas, NV.

[3] Barabási, A.L. (2003). *Linked: How Everything Is Connected to Everything Else and What it*

*Means for Business, Science, and Everyday Life*, Plume Books.

[4] Barabási, A.L. & Albert, Réka. (1999). Emergence of Scaling in Random Networks. *Science* (286): 509-512.

[5] Barabási, A.L., & Bianconi, G. (2001). Competition and Multiscaling in Evolving Networks. *Europhysics Letters* 54 (May 2001): 436-442.

[6] Chung, F., & Graham, R.. (1998). *Erdős on Graphs: His legacy of Unsolved Problems,* (Wellesley, Massachusetts: A K Peters).

[7] Erdős, P. & Rénia, A. (1960). On the Evolution of Random Graphs. *Publ. Math. Inst. Hung. Acad. Sci.* 5: 17-61.

[8] Erdős, P., and Rényi, A., 1959. On random graphs. *Publicationes Mathematicae* (Debrecen) 6, 290–297

[9] Wasserman, S. & Faust, K. (1994). Social *Network Analysis: Methods and Applications,* (Cambridge University Press: 1994).

[10] Granovetter, M. S. (1973). The Strength of Weak Ties. *American Journal of Sociology* 78: 1360-1380.

[11] Hamming, R.W. (1950). Error Detecting and Error Correcting Codes, Bell System Technical Journal 26(2):147-160.

[12] Holland, P.W., & Leinhardt, S. (1971). Transitivity in structural models of small groups. *Comparative Group Studies. 2*: 107-124.

[13] McCulloh, I., Carley, K.M. (n.d.). The Network Probability Matrix: an Alternative to the Exponential Random Graph Model for Longitudinal Data. Unpublished manuscript.

[14] McCulloh, I., Lospinoso, J., and Carley, K.M., (2007). *Social network probability mechanics*. In Proceedings, 12th International Conference on Applied Mathematics, World Scientific Engineering Academy and Society, Cairo, Egypt, pp 319-325

[15] Tichy, N.M., Tushman, M.L., & Fombrun, C. (1979). Social Network Analysis for Organizations. *The Academy of Management Review*, 4(4):507-519

[16] Watts, J. D., & Newman, M. (1999). Renormalization Group Analysis of the Small-World Network Model. *Physics Letters, A*, (263): 341-346.

[17] Watts, J. D., & Strogatz, S. H. (1998). Collective Dynamics of 'Small-World' Networks. *Nature* (393). 440-442.

# Specific Communication Network Measure Distribution Estimation

Daniel P. Baller, Joshua Lospinoso

17 June 2008

## 13th ICCRTS

USMA Network Science Center
http://www.netscience.usma.edu/

# Agenda

- Background & Motivation
- ORA Visualization
- Assumptions
- NPM
- Simulated Networks
- Results
- Future Research

# Background & Motivation

- Simulating random instances of networks is a hot topic in today's Network Science research
  - Erdos-Renyi, Watts-Strogatz, Barabasi-Alberts, NPM
- How do networks arrive at **structure**? How do we explore these structures?

- Many methods of simulating **random networks** exist
- No sound methodology exists for measuring "goodness"

- We propose a methodology for **testing** how well simulations perform under a rigorous **statistical framework**, and execute testing for two data sets under the *Network Probability Matrix* and *Erdos-Renyi*.

# Data Set 1 (Net07)

- Warfighting Simulation run at FT Leavenworth, KS in April 2007.
- 68 Mid Career Army Officers
- 4 day simulated exercise
- Self Reported Communications survey
- Surveys conducted 2 x per day
- 8 Total Data Sets

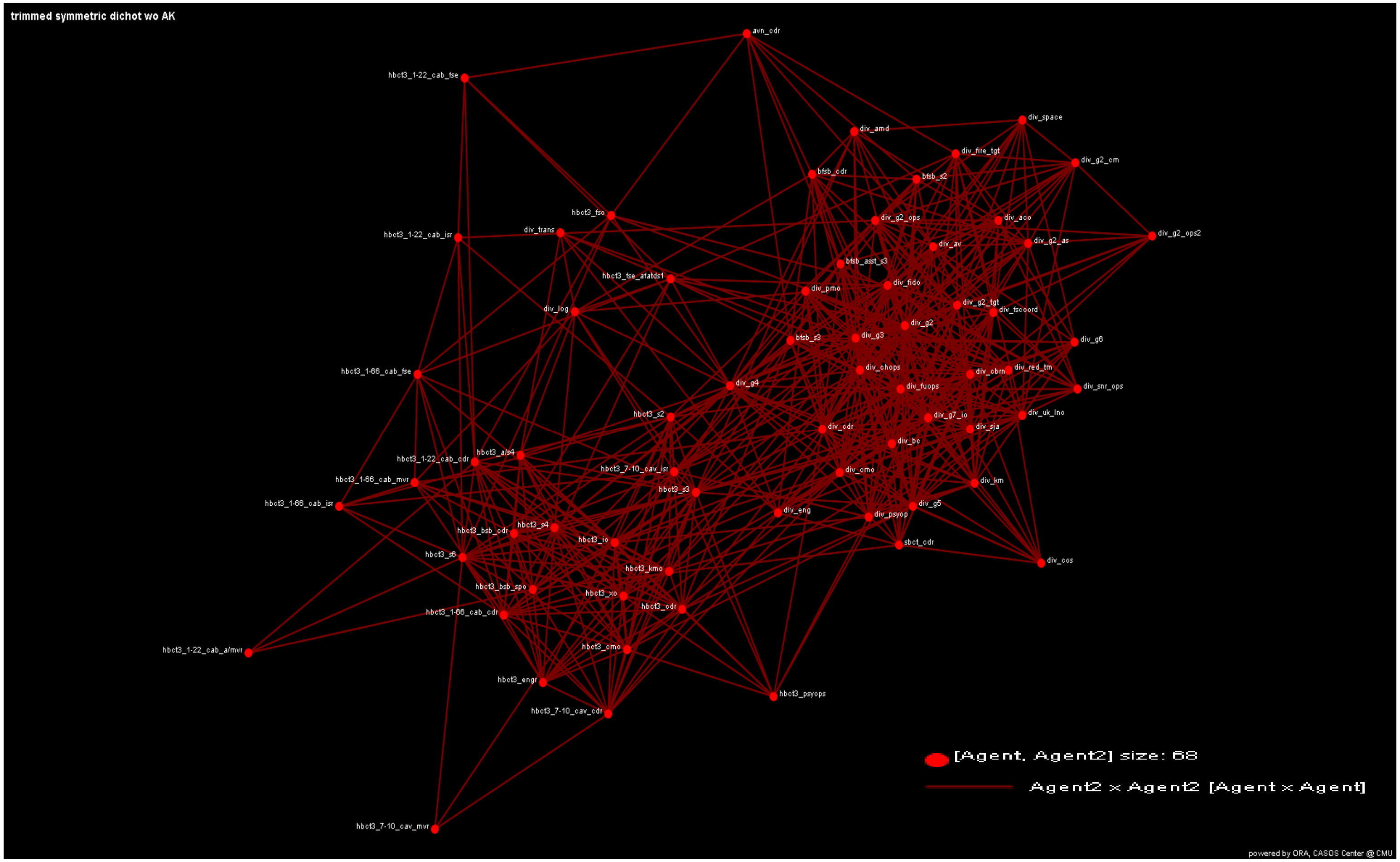
# Data Set 2 (Net05)

- Warfighting Simulation run at FT Leavenworth, KS in 2005.
- 156 Mid Career Army Officers
- 5 day simulated exercise
- Self Reported Communications survey
- Surveys conducted 2 x per day
- 9 Total data sets

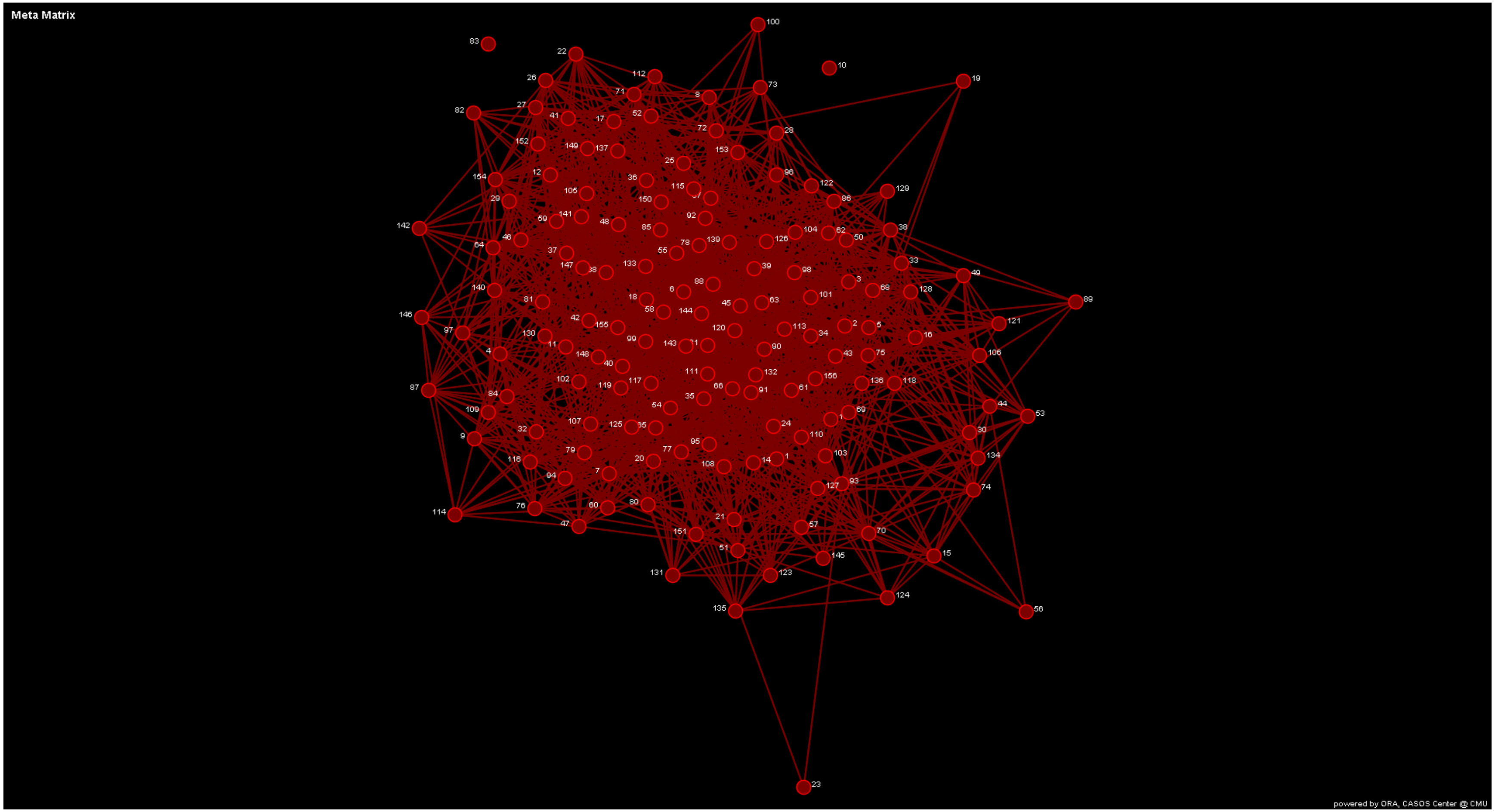
# Data

- Data symmetrized and dichotomized
- Square symmetric matrix
- Over time data compiled by agent

# ORA Visualization (Net07)

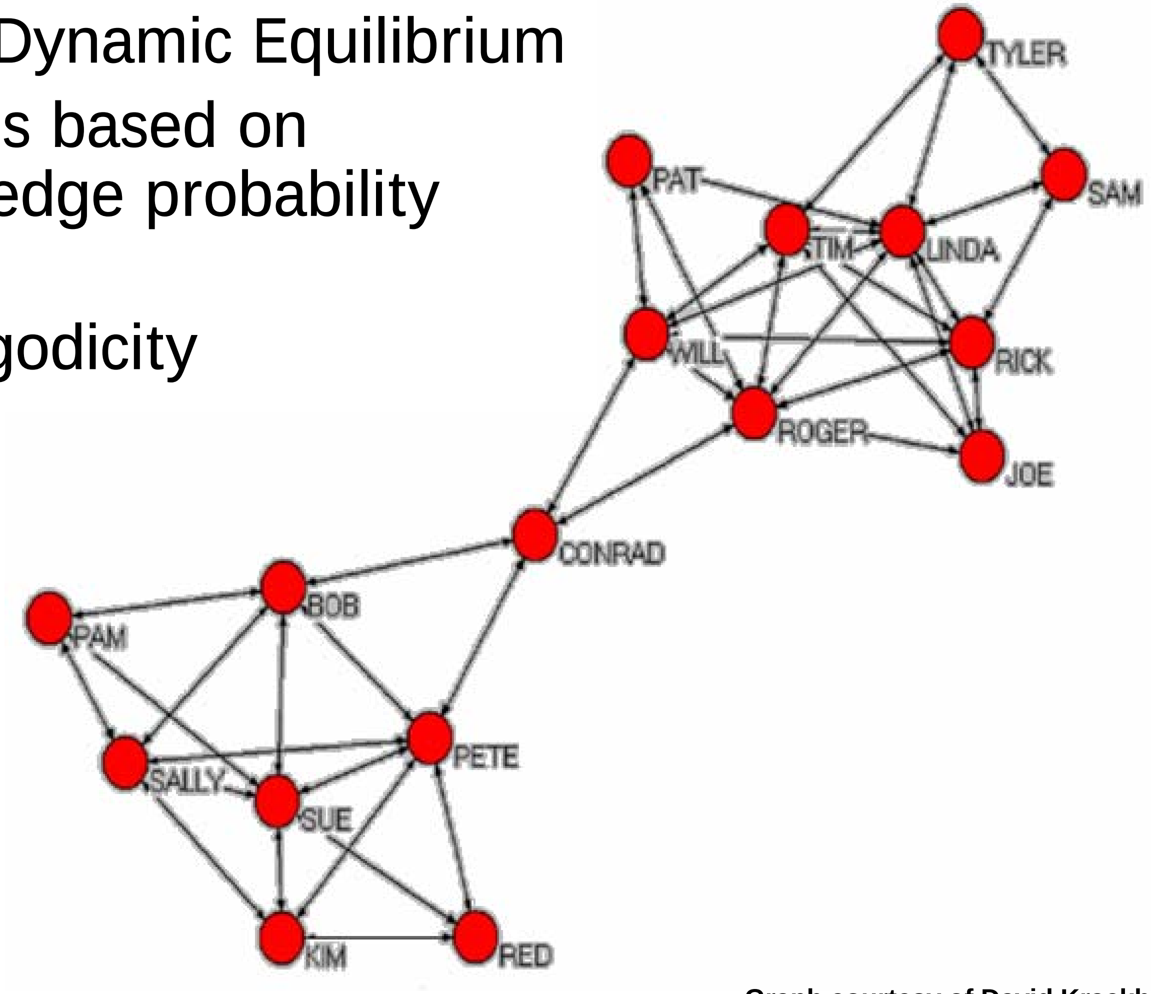


# ORA Visualization (Net05)



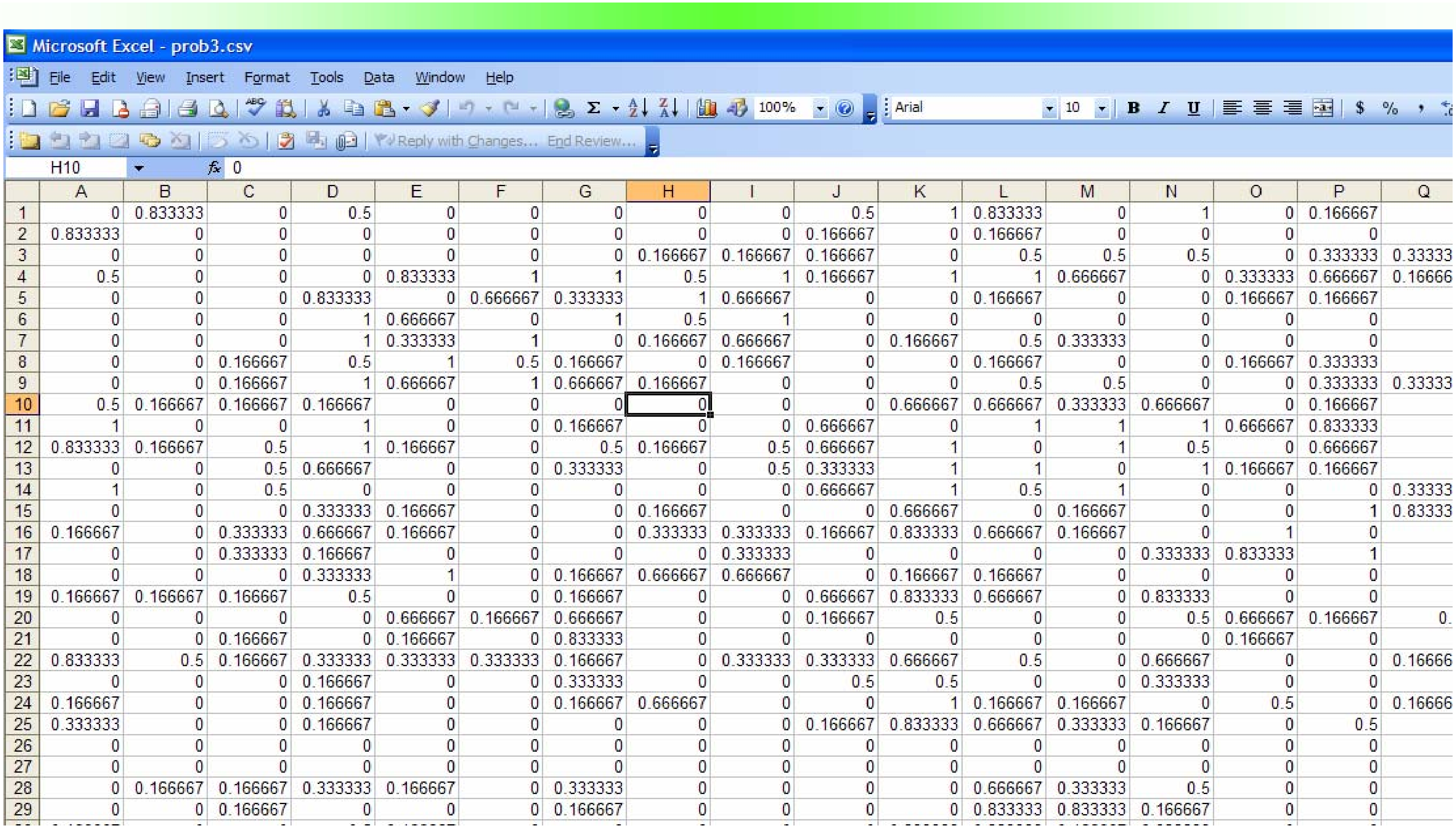
# Assumptions

- Network in Dynamic Equilibrium
- Observations based on Underlying edge probability structure
- Maintain ergodicity



Graph courtesy of David Krackhardt

# Network Probability Matrix



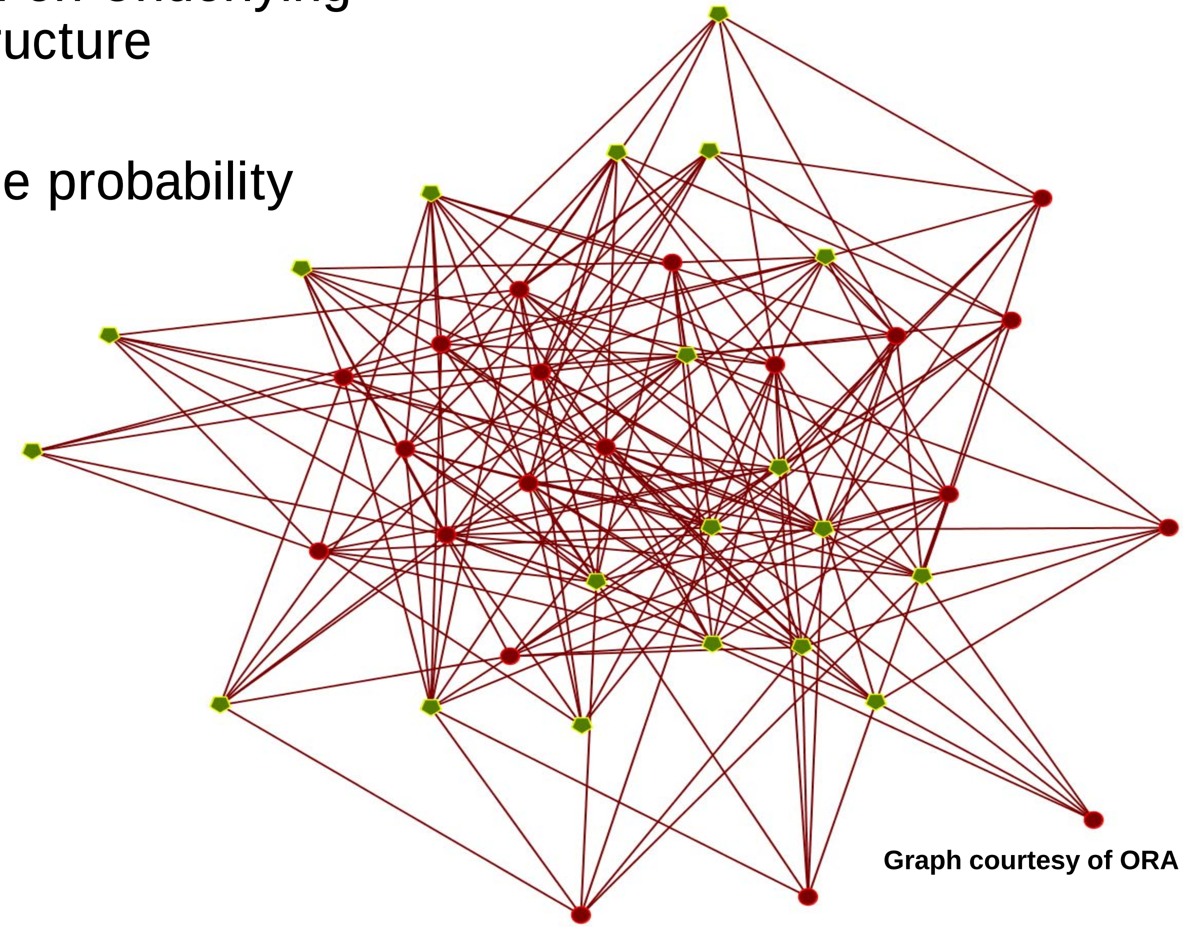
Microsoft Excel - prob3.csv

H10 0

| | A | B | C | D | E | F | G | H | I | J | K | L | M | N | O | P | Q |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 0 | 0.833333 | 0 | 0.5 | 0 | 0 | 0 | 0 | 0 | 0.5 | 1 | 0.833333 | 0 | 1 | 0 | 0.166667 | |
| 2 | 0.833333 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0.166667 | 0 | 0.166667 | 0 | 0 | 0 | 0 | |
| 3 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0.166667 | 0.166667 | 0.166667 | 0 | 0.5 | 0.5 | 0.5 | 0 | 0.333333 | 0.33333 |
| 4 | 0.5 | 0 | 0 | 0 | 0.833333 | 1 | 1 | 0.5 | 1 | 0.166667 | 1 | 1 | 0.666667 | 0 | 0.333333 | 0.666667 | 0.16666 |
| 5 | 0 | 0 | 0 | 0.833333 | 0 | 0.666667 | 0.333333 | 1 | 0.666667 | 0 | 0 | 0.166667 | 0 | 0 | 0.166667 | 0.166667 | |
| 6 | 0 | 0 | 0 | 1 | 0.666667 | 0 | 1 | 0.5 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | |
| 7 | 0 | 0 | 0 | 1 | 0.333333 | 1 | 0 | 0.166667 | 0.666667 | 0 | 0.166667 | 0.5 | 0.333333 | 0 | 0 | 0 | |
| 8 | 0 | 0 | 0.166667 | 0.5 | 1 | 0.5 | 0.166667 | 0 | 0.166667 | 0 | 0 | 0.166667 | 0 | 0 | 0.166667 | 0.333333 | |
| 9 | 0 | 0 | 0.166667 | 1 | 0.666667 | 1 | 0.666667 | 0.166667 | 0 | 0 | 0 | 0.5 | 0.5 | 0 | 0 | 0.333333 | 0.33333 |
| 10 | 0.5 | 0.166667 | 0.166667 | 0.166667 | 0 | 0 | 0 | 0 | 0 | 0 | 0.666667 | 0.666667 | 0.333333 | 0.666667 | 0 | 0.166667 | |
| 11 | 1 | 0 | 0 | 1 | 0 | 0 | 0.166667 | 0 | 0 | 0.666667 | 0 | 1 | 1 | 1 | 0.666667 | 0.833333 | |
| 12 | 0.833333 | 0.166667 | 0.5 | 1 | 0.166667 | 0 | 0.5 | 0.166667 | 0.5 | 0.666667 | 1 | 0 | 1 | 0.5 | 0 | 0.666667 | |
| 13 | 0 | 0 | 0.5 | 0.666667 | 0 | 0 | 0.333333 | 0 | 0.5 | 0.333333 | 1 | 1 | 0 | 1 | 0.166667 | 0.166667 | |
| 14 | 1 | 0 | 0.5 | 0 | 0 | 0 | 0 | 0 | 0 | 0.666667 | 1 | 0.5 | 1 | 0 | 0 | 0 | 0.33333 |
| 15 | 0 | 0 | 0 | 0.333333 | 0.166667 | 0 | 0 | 0.166667 | 0 | 0 | 0.666667 | 0 | 0.166667 | 0 | 0 | 1 | 0.83333 |
| 16 | 0.166667 | 0 | 0.333333 | 0.666667 | 0.166667 | 0 | 0 | 0.333333 | 0.333333 | 0.166667 | 0.833333 | 0.666667 | 0.166667 | 0 | 1 | 0 | |
| 17 | 0 | 0 | 0.333333 | 0.166667 | 0 | 0 | 0 | 0 | 0.333333 | 0 | 0 | 0 | 0 | 0.333333 | 0.833333 | 1 | |
| 18 | 0 | 0 | 0 | 0.333333 | 1 | 0 | 0.166667 | 0.666667 | 0.666667 | 0 | 0.166667 | 0.166667 | 0 | 0 | 0 | 0 | |
| 19 | 0.166667 | 0.166667 | 0.166667 | 0.5 | 0 | 0 | 0.166667 | 0 | 0 | 0.666667 | 0.833333 | 0.666667 | 0 | 0.833333 | 0 | 0 | |
| 20 | 0 | 0 | 0 | 0 | 0.666667 | 0.166667 | 0.666667 | 0 | 0 | 0.166667 | 0.5 | 0 | 0 | 0.5 | 0.666667 | 0.166667 | 0. |
| 21 | 0 | 0 | 0.166667 | 0 | 0.166667 | 0 | 0.833333 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0.166667 | 0 | |
| 22 | 0.833333 | 0.5 | 0.166667 | 0.333333 | 0.333333 | 0.333333 | 0.166667 | 0 | 0.333333 | 0.333333 | 0.666667 | 0.5 | 0 | 0.666667 | 0 | 0 | 0.16666 |
| 23 | 0 | 0 | 0 | 0.166667 | 0 | 0 | 0.333333 | 0 | 0 | 0.5 | 0.5 | 0 | 0 | 0.333333 | 0 | 0 | |
| 24 | 0.166667 | 0 | 0 | 0.166667 | 0 | 0 | 0.166667 | 0.666667 | 0 | 0 | 1 | 0.166667 | 0.166667 | 0 | 0.5 | 0 | 0.16666 |
| 25 | 0.333333 | 0 | 0 | 0.166667 | 0 | 0 | 0 | 0 | 0 | 0.166667 | 0.833333 | 0.666667 | 0.333333 | 0.166667 | 0 | 0.5 | |
| 26 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | |
| 27 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | |
| 28 | 0 | 0.166667 | 0.166667 | 0.333333 | 0.166667 | 0 | 0.333333 | 0 | 0 | 0 | 0 | 0.666667 | 0.333333 | 0.5 | 0 | 0 | |
| 29 | 0 | 0 | 0.166667 | 0 | 0 | 0 | 0.166667 | 0 | 0 | 0 | 0 | 0.833333 | 0.833333 | 0.166667 | 0 | 0 | |

# Erdsös-Réni Random Graph Assumptions

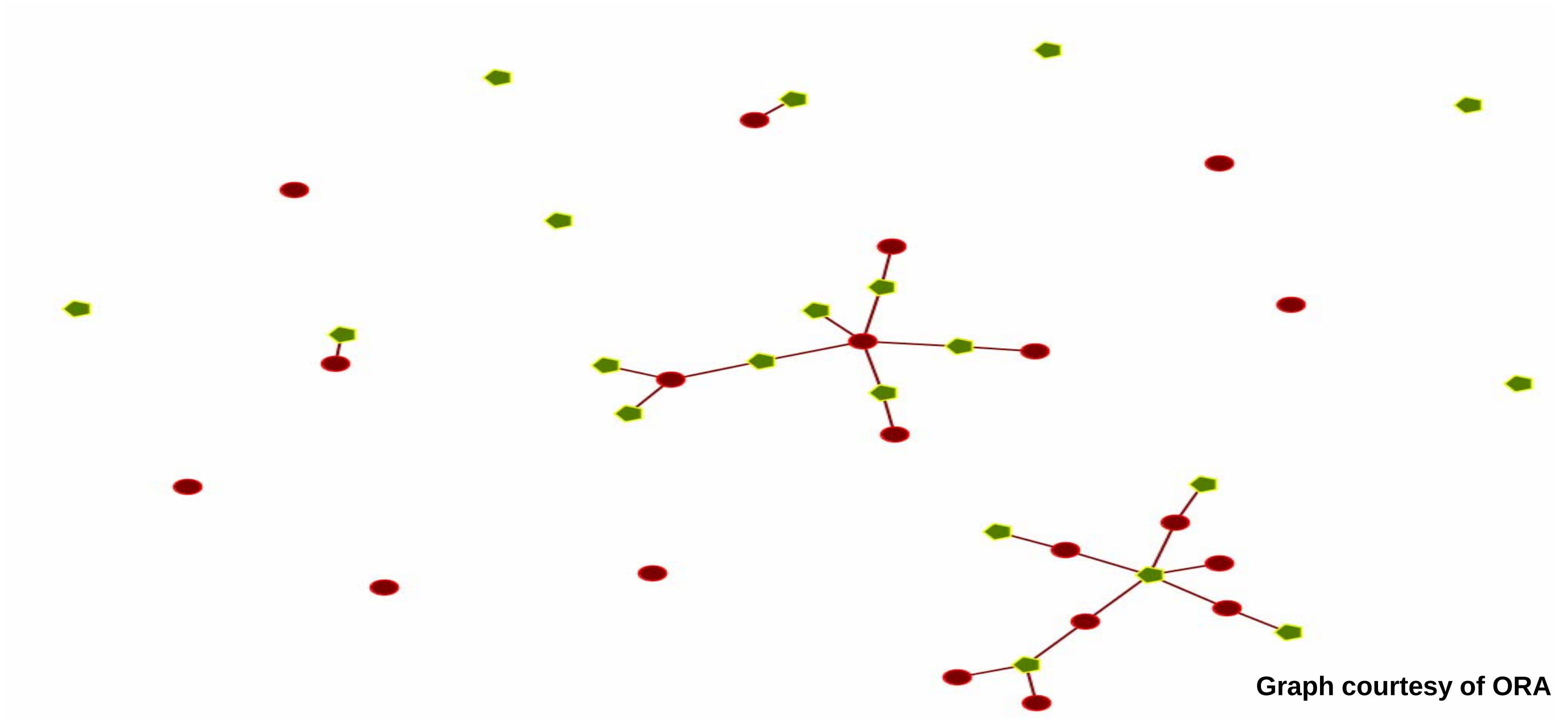
- Network in Dynamic Equilibrium
- Observations based on Underlying edge probability structure
- Maintain ergodicity
- All nodes have same probability distribution.



Graph courtesy of ORA

# Erdsös-Réni Random Graph

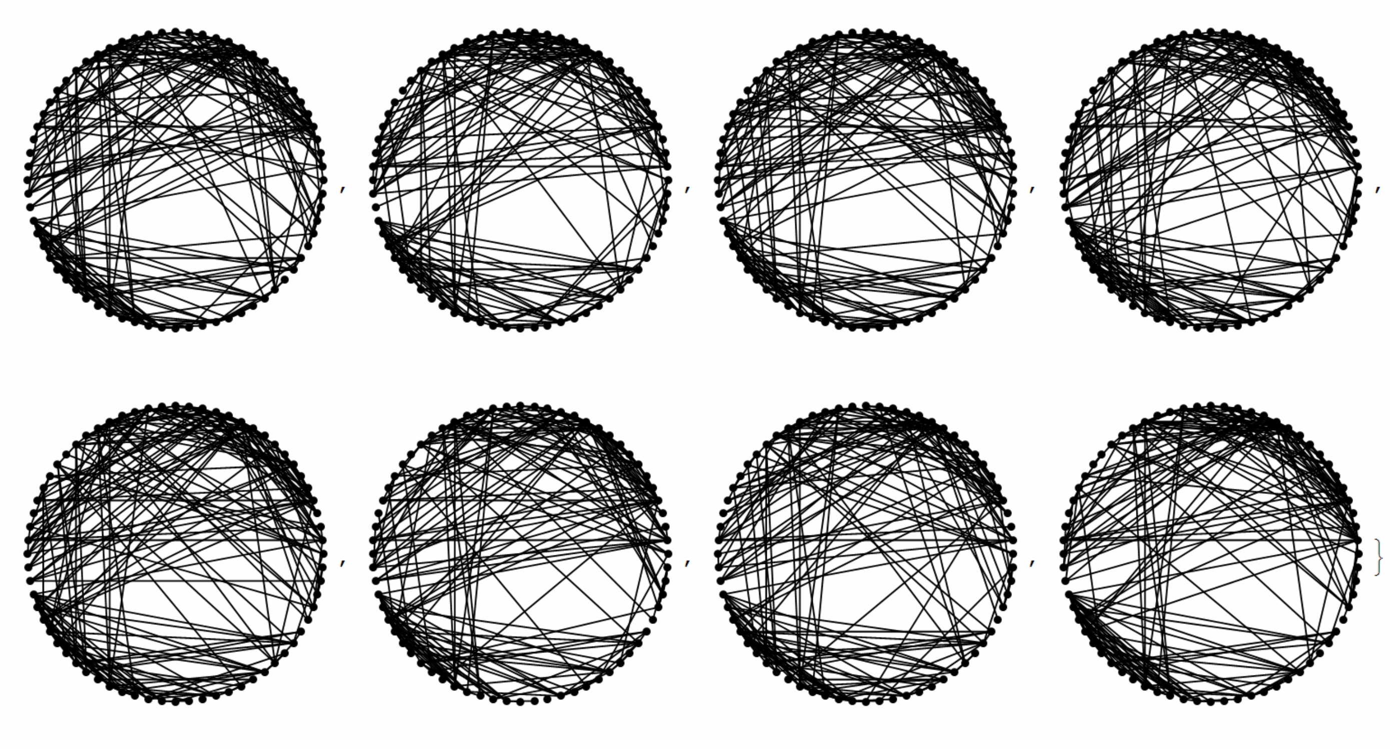
- All edges have the same probability
- Same random Bernoulli trial
- Connectivity threshold $\frac{(1-\varepsilon)\ln(n)}{n}$



Graph courtesy of ORA

# Simulated Network

- Monte Carlo Simulation
- N = 100,000



# The Statistical Test

- We use Hamming Distance (HD) as a measure of similarity between networks.
- We find HD for all combinations of time periods in the empirical data.
- We take all of the simulated graphs, and find HD between them and each time period from empiricals.
- Perform **T-TEST** between these two sets of HDs.
- This test tells us if the simulation (NPM/E-R) does a BETTER JOB at explaining a given time period than the rest of the empirical data does.

# The Statistical Test

1. Create a vector of hamming distances between all possible combinations of **empirical** vectors. Group them by time period.
2. Create a vector of hamming distances between all simulated graphs and **each time period**, grouping the vectors by time period.
3. Perform a T-test between each corresponding vector to answer the following question:

**Does the NPM / E-R**, on average, **more closely match** any particular time period from the empirical data than the rest of the empirical data?

# Results (NET07): NPM

- Comparison of simulated vs. empirical data

| | *M* | *8* | *N* | *60000* | | |
|---|---|---|---|---|---|---|
| **Data Set** | **Mean Hamming Distance to Empirical Networks** | **Standard Deviation of Hamming Distance to Empirical Networks** | **Mean Hamming Distance to Simulated Networks** | **Standard Deviation of Hamming Distance to Simulated Networks** | **t-test** | **p-value** |
| 1 | 409.286 | 38.560 | 358.094 | 12.775 | 3.755 | 0.00 |
| 2 | 365.857 | 18.298 | 320.097 | 12.739 | 7.073 | 0.00 |
| 3 | 365.857 | 29.043 | 320.164 | 12.793 | 4.450 | 0.00 |
| 4 | 377.857 | 38.247 | 330.674 | 12.773 | 3.489 | 0.00 |
| 5 | 375.286 | 36.100 | 328.377 | 12.796 | 3.675 | 0.00 |
| 6 | 349.857 | 38.159 | 306.078 | 12.785 | 3.245 | 0.00 |
| 7 | 373.8571 | 48.45076 | 327.0728 | 12.82622 | 2.731135 | 0.01 |
| 8 | 362.4286 | 55.63529 | 317.1509 | 12.77754 | 2.301849 | 0.02 |

# Results (NET07): E-R

- Comparison of simulated vs. empirical data

| | *M* | *8* | *N* | *60000* | | |
|---|---|---|---|---|---|---|
| **Data Set** | **Mean Hamming Distance to Empirical Networks** | **Standard Deviation of Hamming Distance to Empirical Networks** | **Mean Hamming Distance to Simulated Networks** | **Standard Deviation of Hamming Distance to Simulated Networks** | **t-test** | **p-value** |
| 1 | 409.286 | 38.560 | 1127.379 | 17.41762 | -3.9167 | 0.00 |
| 2 | 365.857 | 18.298 | 1116.303 | 21.54558 | -4.23399 | 0.00 |
| 3 | 365.857 | 29.043 | 1193.895 | 18.60198 | -3.73844 | 0.00 |
| 4 | 377.857 | 38.247 | 1252.086 | 16.82216 | -4.40049 | 0.00 |
| 5 | 375.286 | 36.100 | 1169.254 | 18.88182 | -3.64695 | 0.00 |
| 6 | 349.857 | 38.159 | 1209.797 | 17.59757 | -3.60082 | 0.00 |
| 7 | 373.8571 | 48.45076 | 1110.78 | 17.31786 | -3.44968 | 0.00 |
| 8 | 362.4286 | 55.63529 | 1192.288 | 17.44347 | -3.461 | 0.00 |

# Results (NET05): NPM

| | *M* | *9* | *N* | *60000* | | |
|---|---|---|---|---|---|---|
| **Data Set** | **Mean Hamming Distance to Empirical Networks** | **Standard Deviation of Hamming Distance to Empirical Networks** | **Mean Hamming Distance to Simulated Networks** | **Standard Deviation of Hamming Distance to Simulated Networks** | **t-test** | **p-value** |
| 1 | 1445 | 84.774 | 1284.338 | 23.747 | 3.467 | 0.001 |
| 2 | 1394.75 | 67.487 | 1239.647 | 23.703 | 3.765 | 0.000 |
| 3 | 1296.125 | 85.436 | 1151.946 | 23.671 | 3.287 | 0.001 |
| 4 | 1315.875 | 153.533 | 1169.665 | 23.718 | 2.421 | 0.015 |
| 5 | 1191.25 | 112.324 | 1058.99 | 23.667 | 2.732 | 0.006 |
| 6 | 1204.875 | 207.944 | 1071.116 | 23.623 | 1.912 | 0.056 |
| 7 | 1167.375 | 190.431 | 1037.713 | 23.695 | 1.98 | 0.048 |
| 8 | 1159.625 | 204.465 | 1030.815 | 23.732 | 1.888 | 0.059 |
| 9 | 1170.125 | 195.266 | 1040.142 | 23.618 | 1.953 | 0.051 |

# Results (NET05): E-R

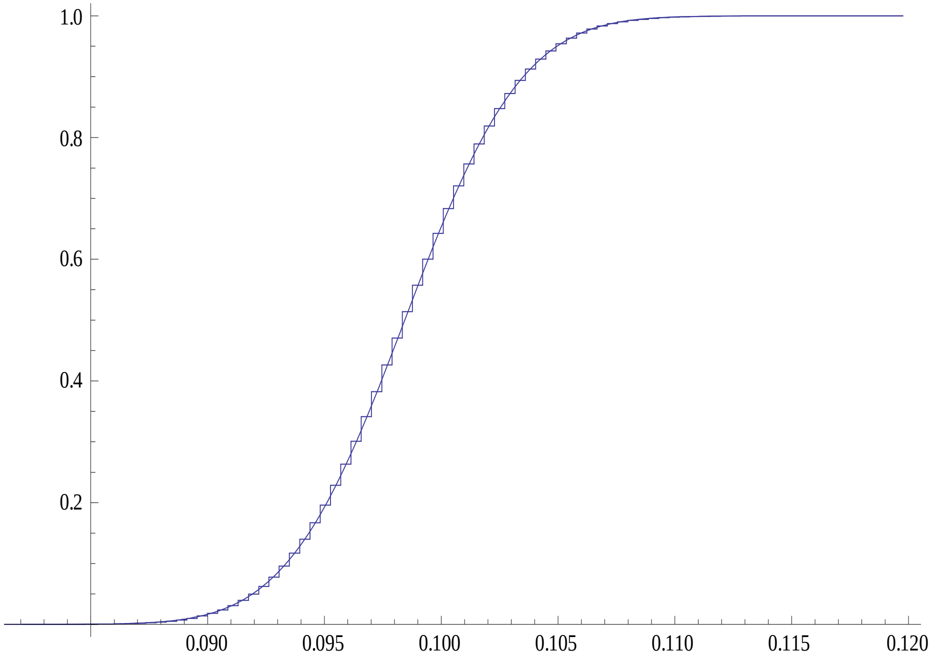
| | *M* | *9* | *N* | *60000* | | |
|---|---|---|---|---|---|---|
| **Data Set** | **Mean Hamming Distance to Empirical Networks** | **Standard Deviation of Hamming Distance to Empirical Networks** | **Mean Hamming Distance to Simulated Networks** | **Standard Deviation of Hamming Distance to Simulated Networks** | **t-test** | **p-value** |
| 1 | 1445.000 | 84.774 | 2253.82 | 34.26138 | -6.8034 | 0.00 |
| 2 | 1394.750 | 67.487 | 2232.07 | 41.48661 | -7.46798 | 0.00 |
| 3 | 1296.125 | 85.436 | 2385.99 | 35.58944 | -6.47687 | 0.00 |
| 4 | 1315.875 | 153.533 | 2503.9 | 32.87007 | -7.80098 | 0.00 |
| 5 | 1191.250 | 112.324 | 2336.64 | 36.9779 | -6.2939 | 0.00 |
| 6 | 1204.875 | 207.944 | 2419.19 | 34.87729 | -6.20163 | 0.00 |
| 7 | 1167.375 | 190.431 | 2219.81 | 33.75171 | -5.89936 | 0.00 |
| 8 | 1159.625 | 204.465 | 2383.33 | 33.89981 | -5.99199 | 0.00 |
| 9 | 1170.125 | 195.266 | 2453.82 | 36.2168 | -7.1034 | 0.00 |

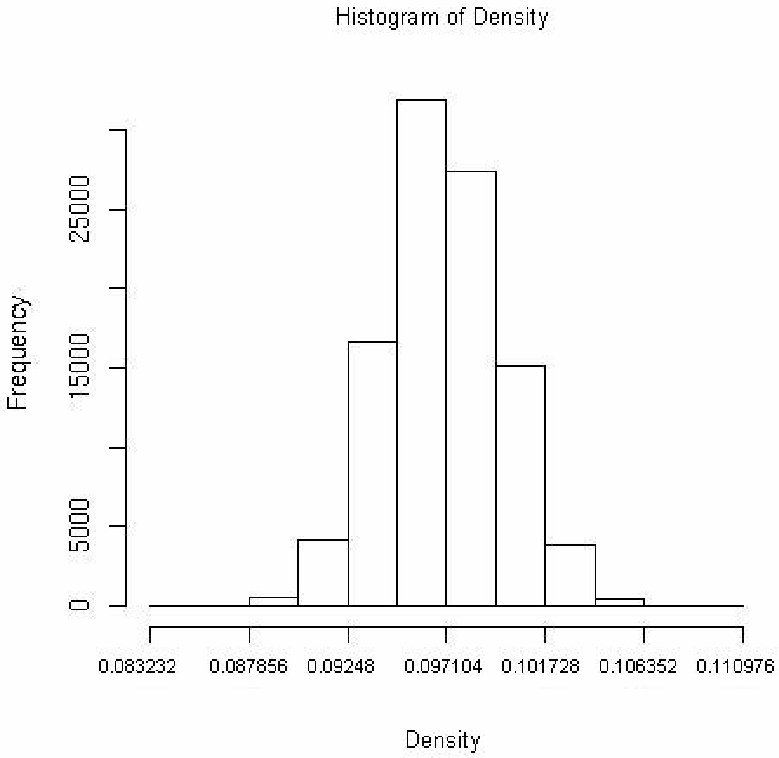
# Significance

- NPM performs well, E-R does not.
  - Why? (Net-07 Clustering)
- Since the NPM does a good job of representing the laws which govern the network, we can use simulation to:
  - Explore large numbers of “instances” of the graphs
  - Create distributions of network and agent-level measures
- With a validated simulation, we facilitate further statistical analysis of the network and its measures!
  - Statistical Process Control: When has the network undergone a significant change?
  - Percolation: What is the likelihood that a rumor/ideology/belief spreads throughout the network?

# Results (NET07)

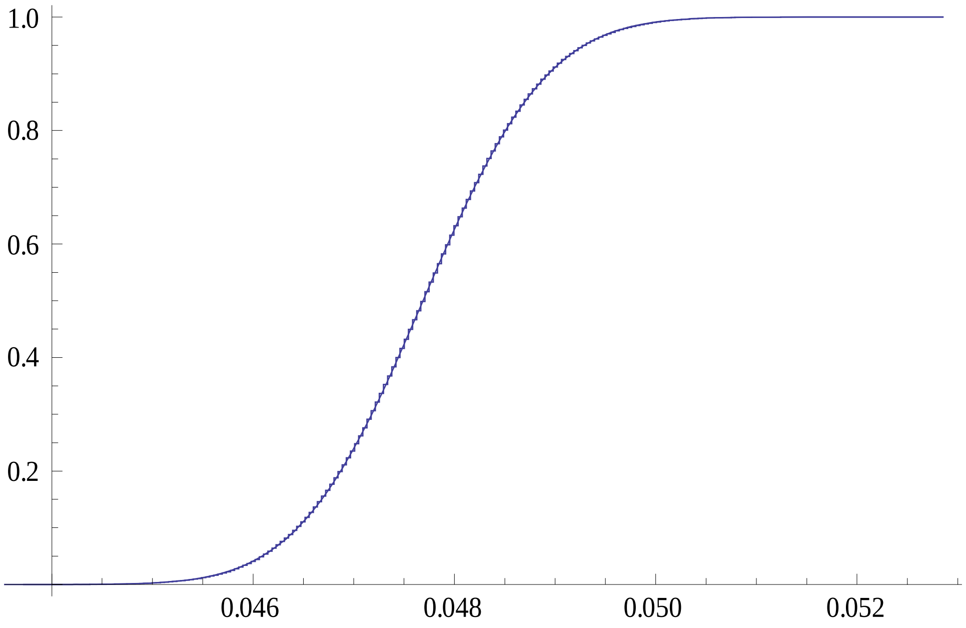
- Fit a normal distribution to densities

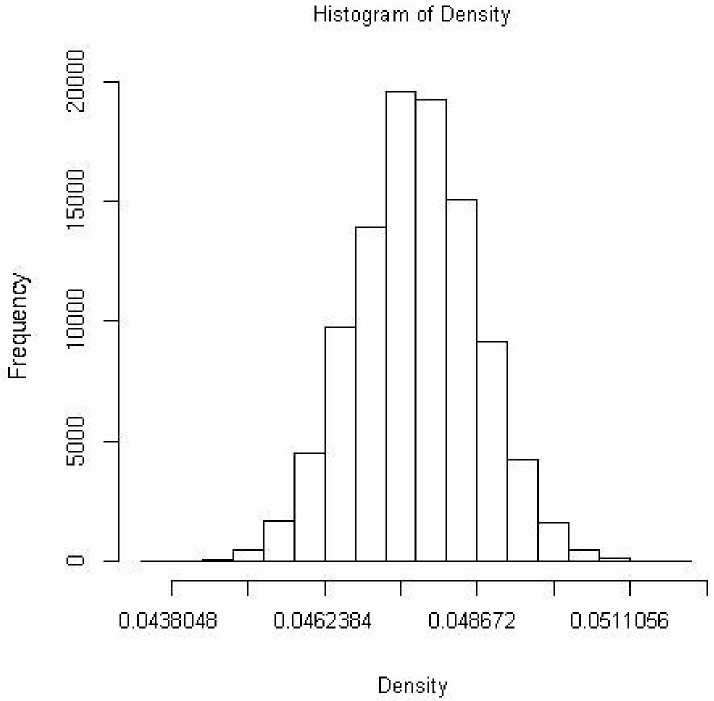




# Results (NET05)

- Fit a normal distribution to densities





# Conclusion

- When highly complex systems are being simulated, and empirical data is available, we can use this methodology to test whether our simulation is *at least as close* to each time series in a data set as the rest of the time periods are.
  - Which model (Erdos, Watts, Barabasi, NPM) most accurately describes the empirical data?

- The “simple case” of the NPM is shown to be a viable explanation of social networks.

# Acknowledgements

- U.S. Army Research Institute
- U.S. Military Academy Network Science Center
- U.S. Army Research Labs
- Dr. Kathleen Carley
- COL Steve Horton
- LTC John Graham
- MAJ Tony Johnson
- MAJ Ian McCulloh
- Dr. Craig Schrieber